

# Principles of Study Design in Environmental Epidemiology

Hal Morgenstern<sup>1\*</sup> and Duncan Thomas<sup>2</sup>

<sup>1</sup>Department of Epidemiology, University of California at Los Angeles, School of Public Health, Los Angeles, CA 90024-1772; <sup>2</sup>Department of Preventive Medicine, University of Southern California, School of Medicine, Los Angeles, CA 90033-9987

This paper discusses the principles of study design and related methodologic issues in environmental epidemiology. Emphasis is given to studies aimed at evaluating causal hypotheses regarding exposures to suspected health hazards. Following background sections on the quantitative objectives and methods of population-based research, we present the major types of observational designs used in environmental epidemiology: first, the three basic designs involving the individual as the unit of analysis (i.e., cohort, cross-sectional, and case-control studies) and a brief discussion of genetic studies for assessing gene-environment interactions; second, various ecologic designs involving the group or region as the unit of analysis. Ecologic designs are given special emphasis in this paper because of our lack of resources or inability to accurately measure environmental exposures in large numbers of individuals. The paper concludes with a section highlighting current design issues in environmental epidemiology and several recommendations for future work. — *Environ Health Perspect* 101(Suppl 4):23-38 (1993).

**Key Words:** Study design, epidemiologic methods, environmental health, ecologic studies, aggregate studies, causal inference

## Introduction

The purpose of this article is to discuss the principles of study design and related methodologic issues in environmental epidemiology. The focus is on studies aimed at evaluating causal hypotheses regarding exposures to suspected health hazards. Because the intended audience for this document includes scientists without formal training in epidemiology, parts of this article highlight basic principles of epidemiologic research. Nevertheless, we also try to summarize comprehensively the current state of the art and make recommendations for future developments in study design. For more extensive treatment of general research principles and methods in epidemiology, the interested reader should consult available textbooks in this area (1-6). More detailed examples of applications in environmental epidemiology may be found in several other books, such as those edited by Leaverton (7), Chiazze et al. (8), Goldsmith (9), and Kopfler and Craun (10).

## Population Parameters

The major quantitative objectives of most epidemiologic studies are to estimate two types of population parameters: the frequency of disease occurrence in particular populations and

the effect of a given exposure on disease occurrence in a particular population.

Measures of disease frequency involve the occurrence of new cases or deaths (incidence/mortality) or the presence of existing cases (prevalence). In both applications, the number of cases is expressed relative to the size of the population from which the cases are identified. With incidence measures, this denominator is the (base) population at risk (i.e., individuals who are eligible to become cases). Thus, the base population of a study (or study base) is the group of all individuals who, if they developed the disease, would become cases in the study (3,11,12).

Disease incidence, which is central to the process of causal inference, can be expressed as a cumulative measure (risk) or as a person-time measure (rate). The cumulative incidence (incidence proportion) or average risk in a base population is the probability of someone in that population developing the disease during a specified period, conditional on not dying first from another disease (13). The term cumulative incidence or cumulative incidence rate also is defined somewhat differently as the integral over the follow-up period of the hazard (rate) function (14). The incidence rate or instantaneous risk (hazard) is the limit of the average risk for a given period, per unit of time, as the duration of the period approaches zero. The average rate (incidence density) for a given period is estimated as the number of incident events divided by the amount of person-time experienced by the base popula-

tion. For example, a rate of 0.001/year means that we would expect one new case to occur for every 1000 person-years of follow-up (e.g., 100 disease-free people followed for an average of 10 years).

Although there are many quantitative methods for expressing the magnitude of a statistical association between two variables (e.g., exposure status and disease occurrence), we are usually interested in a special class of such measures that reflect the net effect of the exposure on disease occurrence (i.e., causal parameters). In general, a causal parameter for a target population is a hypothetical contrast—in the form of a difference or ratio—between what the frequency of disease would be if everyone were exposed (at a given level) to what the frequency would be if everyone were unexposed (often called the reference level) (15). When this difference for a specific exposure is not zero (the ratio is not one), we say that the exposure is a risk factor for that disease in the target population. In practice, we estimate causal parameters indirectly by comparing disease frequency for an exposed group with disease frequency for an unexposed group. Epidemiologists typically estimate the risk or rate ratio (often called the relative risk) by comparing the exposed population with an unexposed population. The key assumption of this statistical approach is that the risk or rate observed for the unexposed group is the same (within confounder strata) as the risk or rate that would have been observed in the exposed group if that group had not been exposed

This manuscript was prepared as part of the Environmental Epidemiology Planning Project of the Health Effects Institute, September 1990–September 1992.

\*Author to whom correspondence should be addressed.

This work was funded by the Health Effects Institute in Cambridge, MA. The authors would like to thank Dr. John Tukey, Dr. Sander Greenland, and other members of the HEI Methodology Working Group for their helpful comments.

(16). Thus, the (true) risk ratio may be interpreted as a causal parameter, which is the number of cases actually occurring in the exposed (target) population divided by the number of cases that would have occurred in the absence of exposure.

Certain measures of association, such as correlation coefficients and standardized regression coefficients, do not, in general, reflect any causal parameters. The reason is that the magnitude of these measures depends in part on the relative variances of the exposure and disease variables, which are influenced by the sampling strategy (i.e., noncausal parameters) (17,18). Another measure of association, the odds ratio, is used in certain types of epidemiologic studies (case-control designs) to estimate the risk or rate ratio indirectly when we cannot first estimate the incidence rate or risk in the exposed and unexposed populations (1-6,19,20).

### Problems in Environmental Epidemiology

There are several general problems in environmental epidemiology that tend to limit causal inference and, therefore, shape design decisions.

**Long Latent Periods.** The interval between first exposure to an environmental risk factor (or the start of causal action of this factor) and disease detection (or symptom onset) may be many years or even decades. Such long latent periods are partly due to limitations of medical technology and incomplete surveillance for detecting disease; yet they are also due to a prolonged induction period in which years are needed for the disease process to begin (5). The term latent period also is used more specifically to indicate the hypothetical interval between disease initiation and detection (5). Refer also to Armenian and Lilienfeld (21) who discuss alternative definitions of latency. Unfortunately, long latent periods produce important practical constraints on our ability to estimate exposure effects. The investigator must either observe subjects for many years or rely on retrospective (historical) measurement of key variables. The latter alternative may be infeasible for certain types of exposures or in certain populations. Even when feasible, however, retrospective measurement usually increases the amount of error with which exposures are measured (see below). Furthermore, the level of most environmental exposures and many extraneous risk factors changes appreciably or unpredictably over time; long latent periods, therefore, seriously complicate our ability to estimate effects (22).

**Errors of Exposure Measurement.** A major challenge in environmental epidemi-

ology is to measure accurately each individual's exposure to hypothesized risk factors (i.e., the biologically relevant dose [Thomas and Hatch, this issue]). This task is made very difficult by the lack of information about environmental sources of emission, the complex pattern of most long-term exposures, the individual's ignorance of previous opportunities for exposure, the lack of good biological indicators of exposure level, and the lack of sufficient resources to collect individual exposure data on large populations. The consequences of exposure mis-measurement are probable bias in the estimation of effect (see "Sources of Epidemiologic Bias") and possible loss of precision and power with which effects are estimated and tested (23,24). The problem and issues of exposure measurement are discussed more thoroughly by Hatch and Thomas in this issue.

**Rare Diseases, Low-Level Exposures, and Small Effects.** In most epidemiologic studies of environmental hazards, statistical objectives may be further compromised by the infrequent occurrence of the disease or outcome of interest, by the low prevalence or levels of environmental exposures in the general population, and by the search for small effects (for which the true rate ratio is between 0.5 and 2). A critical consequence of these features is usually substantial loss of precision and power with which effects are estimated and tested. In addition, it becomes more difficult for the investigator to separate the effect of the exposure of interest from the distorting effects of extraneous factors. Causal inference can then be seriously compromised.

### Research Objectives and Strategies

Given the above problems, epidemiologists must carefully plan their studies, analyze their data, and interpret their findings. Inaccurate results reflect both random errors of estimation (chance) and systematic errors or bias. An epidemiologically unbiased or valid estimate of a causal parameter is one that is expected to represent perfectly (aside from chance) the true value of the parameter in the base population.

#### Sources of Epidemiologic Bias

A common framework for describing the validity of epidemiologic research is to consider three sources of bias in the estimation of effect: selection bias, information bias, and confounding (2). Despite the practical attractiveness of this framework, the three types of bias are not entirely separate concepts. The amount of confounding, for

example, can depend on how subjects are selected.

**Selection Bias.** Selection bias means that the way in which subjects are selected into the study population or into the analysis (due to lost subjects or missing data) distorts the effect estimate. In general, this problem occurs when either disease status or exposure status influences the selection of subjects to a different extent in the groups being compared. Selection bias is most likely to be problematic when the investigator does not identify the base population from which study cases arose.

**Information Bias.** Information bias means that the nature or quality of measurement or data collection distorts the effect estimate. The primary source of information bias is error in measuring one or more variables. When exposure status or disease status is misclassified, bias usually occurs. If the probabilities of misclassification of each variable are the same for each category of the other variable (nondifferential misclassification) and if the errors for different variables are independent, the estimate of effect is usually biased toward the null value (indicating no effect). Possible exceptions to this principle of nondifferential misclassification leading to conservative effect estimates arise when the misclassified exposure variable is categorized into more than two groups (25). In other situations involving differential misclassification (unequal misclassification probabilities) or correlated measurement errors, the effect estimate may be biased in either direction. In many studies, therefore, the magnitude of misclassification bias is difficult to predict, especially when other biases are operating.

**Confounding.** Confounding refers to a lack of comparability between exposure groups (e.g., exposed versus unexposed) such that disease risk would be different even if the exposure were absent or the same in both populations (16). Thus, confounding is epidemiologic bias in the estimation of a causal parameter (see "Population Parameters"). Because there is no empirical method for directly observing the presence or magnitude of confounding, in practice we attempt to identify and control for manifestations of confounding. This is done by searching for differences between exposure groups in the distribution of extraneous risk factors for the disease, which are called confounders. Thus, a confounder is a risk factor (or proxy) that is associated with exposure status in the base population. A covariate meeting these criteria is not a confounder, however, if its association with the exposure is due entirely to

the effect of the exposure on the covariate; for example, the covariate might be an intermediate variable in the causal pathway between the exposure and disease. If the exposure and covariate are time-dependent variables, it is possible for that covariate to be both a confounder and an intermediate variable (see "Cohort Study").

### The Need for Covariate Data

In addition to the exposure of interest, there is the need in virtually all epidemiologic studies to collect data on other known or possible risk factors for the disease. These covariates may be relevant to the exposure effect in three ways: *a*) as confounders, *b*) as intermediate variables, and *c*) as effect modifiers.

The effects of confounders must be controlled or removed analytically to obtain unbiased estimates of causal parameters. This control is usually achieved through stratification or model fitting. The assessment and control of intermediate variables can elucidate causal mechanisms that explain exposure effects (26). This approach often leads to new etiologic hypotheses and new intervention strategies for disease prevention.

When the exposure–effect measure varies across categories or levels of another factor, we call the second factor an effect modifier; this statistical phenomenon is called effect modification or an interaction effect. The assessment of effect modification is model-dependent, meaning that it depends on what (causal) parameter is used to measure the effect (2–6). For example, an extraneous risk factor that does not modify the risk ratio for the exposure will modify the risk difference. The assessment of effect modification is important for properly specifying the predictors in statistical models (2,14), for making inferences about possible biological (causal) interactions between exposures (e.g., synergy) (5), and for generalizing one's results to other populations (see "Cohort Study").

### Types of Research

There are three general design strategies for conducting population research: *a*) experiments in which the investigators randomly assign (randomize) subjects to two or more treatment (exposure) groups; *b*) quasi-experiments in which the investigators make the assignments to treatment groups nonrandomly; and *c*) observational studies in which the investigators simply observe exposure (treatment) status in subjects without assignment (2). Although some epidemiologists classify the first two types as intervention studies, obser-

vation studies might also involve the evaluation of an intervention that was not implemented or controlled by the investigators. Social scientists often use the term quasi-experiment to mean any type of nonrandomized study (27).

**Experiments.** In a simple experiment, there are usually two treatment groups. One group is assigned to receive the new experimental intervention and the other (control) group is assigned to receive no intervention, a sham intervention (placebo), or another available intervention. Simple randomization of individuals to treatment groups implies that all possible allocation schemes of assigned subjects are equally likely (28). Following randomization, the investigator follows subjects for subsequent disease occurrence or change in outcome status. A comparison of risks between treatment groups provides an estimate of a causal parameter reflecting the treatment effect.

Because experiments are best suited ethically and practically to the study of health benefits, not hazards, experiments in environmental epidemiology would usually be limited to the study of preventive interventions. Furthermore, it is generally impossible or infeasible to randomize subjects individually. The only practical alternative, therefore, is to randomize by group, where the group might be a city, school, work site, etc. (29). The major limitation of group randomization is some within-group dependence (correlation) of the outcome variable, which reduces precision and power (30,31). Thus, the effective sample size falls between the number of randomized groups and the total number of subjects (see Prentice and Thomas, this issue).

As an example, consider the hypothesis that the intake of fluoride ions in drinking water has a protective effect on the occurrence of dental caries in children. An experiment might be conducted by randomly assigning many water districts (each with one fluoride-deficient water supply without treatment) either to implement sodium fluoride treatment under the control of the investigators or to continue its current policy of no treatment for the duration of follow-up. Assuming the hypothesis were true, we would expect the subsequent incidence rate of dental caries to be lower in the treated districts than in the untreated districts.

Randomization insures a valid comparison of subjects according to intended treatment, i.e., assigned treatment, but not according to treatment actually received (16,28). That is, randomization of a sufficient number of units (subjects or groups) provides some assurance that the assigned

treatment groups are comparable with respect to inherent risk. This does not imply that there can be no confounding in a comparison of randomly assigned groups. Even with perfect adherence to treatment assignments and no loss to follow-up, assigned groups might have, by chance, different hypothetical risks in the absence of treatment. Nevertheless, such confounding, if it exists, is equally likely to be positive or negative; conventional confidence-interval estimates and *p* values reflect the possibility of this bias, which becomes smaller as the (effective) sample size increases (28). This protection against confounding afforded by randomization, however, does not apply to lack of adherence or loss to follow-up, both of which usually do not occur randomly. Furthermore, if some subjects cross over between treatments (e.g., residents of a fluoridated district obtain their water from non-fluoridated districts), a comparison of assigned groups will underestimate the true treatment effect even when the crossover is random (32). A comparison of compliers with noncompliers, on the other hand, is essentially observational and therefore prone to bias.

**Quasi-Experiments.** A quasi-experiment may be done similarly to an experiment by comparing two or more nonrandomized groups, or it may be done by comparing one or more groups over time, before versus after the intervention is initiated in at least one group. With the latter approach, the composition of each group may change over time so that subjects observed before the intervention are not the same subjects observed after the intervention.

Returning to the fluoride hypothesis, a quasi-experiment was done in the 1940s and 1950s by comparing two similar, nearby cities in New York State, both of which lacked fluoride treatment before 1945. Newburgh started sodium fluoride treatment in 1945 and continued throughout the 10-yr postintervention follow-up period; Kingston continued to use its fluoride-deficient water without treatment (33). The investigators found that the rate of decayed, missing, or filled (DMF) teeth in children, ages 6 to 12, decreased by almost 50% in Newburgh but increased slightly in Kingston.

Because subjects were not individually randomized in this study, it is possible that children in the treated group differed from children in the comparison group with respect to other risk factors for tooth decay, such as diet. Thus, the investigators' comparisons might have been confounded. Note, however, that randomization by city would not have reduced this possible bias

in the Newburgh–Kingston study, because the two assigned treatment groups would be equally noncomparable regardless of which city was assigned fluoride treatment.

**Observational Studies.** Unlike experiments and quasi-experiments, observational studies are commonly used to estimate the effects of exposures hypothesized to be harmful, fixed attributes (e.g., race and genotype), characteristics, behaviors or exposures over which the investigator has little or no control (e.g., weight, depression, and sunlight exposure), and other exposures for which manipulation or randomization would be unethical or infeasible. Observational studies are often conducted with secondary or retrospective data (instead of primary prospective data) and/or without following individual subjects for change in disease status. For example, the fluoride hypothesis could be tested by comparing the prevalence of decayed, missing, or filled teeth in children who live in areas supplied by fluoridated water with the corresponding prevalence in children who live in areas supplied by nonfluoridated water. Although such a study would be less expensive and easier to conduct than would the previous examples, there are additional methodologic problems that could lead to bias or misinterpretations.

The remainder of this article is devoted to an elaboration of observational study designs. In “Basic Observational Designs,” we cover the basic designs in which data on disease status, exposure status, and all covariates are collected at the individual level; that is, the unit of analysis is the individual (or body part, such as the tooth or eye). In “Ecological Designs,” we cover designs in which the unit of analysis is a group of individuals, such that information is missing on the joint distributions of key variables at the individual level.

## Basic Observational Designs

Frequently, hypotheses about environmental risk factors for disease are derived from animal studies, clinical observations, reports of disease clusters, descriptive findings from population surveillance systems, and various types of exploratory studies (e.g., case series, mapping studies, and migrant studies). Formal testing of these hypotheses most often proceeds by conducting observational studies of the types described in this section.

Basic designs in epidemiology may be classified according to two dimensions: type of study population and type of sampling scheme (34). First, the study population is longitudinal, involving the detection of incident events during a follow-up period; or it is cross-sectional, involving the detection of prevalent

cases at one time. Second, the sampling strategy involves complete selection of the entire population from which study cases are identified, or it involves incomplete or case-control sampling of a fraction (<100%) of the non-cases in the population from which study cases are identified. Case-control sampling, therefore, implies stratification on disease status in the selection process. Combining these two dimensions results in four basic designs: longitudinal studies of a complete population (cohort studies); cross-sectional studies of a complete population (cross-sectional studies); longitudinal studies with case-control sampling (case-control studies with incident cases); and cross-sectional studies with case-control sampling (case-control studies with prevalent cases). In addition to these basic designs, we also discuss new developments in genetic studies for assessing gene–environment interactions (see “Genetic Studies”).

## Cohort Study

A cohort or follow-up study is a longitudinal design of a specified population in which exposure status is measured for all subjects at the start of follow-up (baseline) and possibly during follow-up. The entire study population—typically persons who are free of the index disease at baseline—are followed for detection of all incident cases or deaths of interest. Thus, the base population in this design is identical to the study population.

Cohort studies may be entirely prospective, meaning exposure status and disease occurrence are ascertained for the period during which the study is conducted, or they may be entirely retrospective (historical), meaning exposure status and disease occurrence are ascertained for a period before the study begins. Retrospective data are usually obtained from the subject's recall of past events or from abstracted records. Many cohort studies combine both data-collection procedures; e.g., the follow-up period for detecting the disease starts before the study and continues throughout the study period. Although retrospective studies are generally much less expensive and time-consuming, prospective studies can be designed to collect more appropriate, complete, and accurate data.

**Example.** Suppose we want to estimate the possible effect of prenatal exposure to passive smoke (not maternal smoking) on the risk of lower respiratory disease during the first 3 years of life. We might identify a large group of nonsmoking pregnant women and interview them just before delivery about their exposure to passive smoke during pregnancy and about other

risk factors for the disease. The assessment of passive smoking would involve measuring exposure at home, work, and elsewhere with an attempt to quantify the number of smokers, cigarettes, and/or exposure time for each woman by trimester. Then each neonate would be followed by periodic examinations and parental reports of symptoms to his or her third birthday. By establishing a standard set of criteria for diagnosing new cases of lower respiratory disease and by categorizing the passive-smoke exposure into two or more categories, we can compare the 3-year risk of disease by exposure group. In this hypothetical example, the experience of each subject contributes to a single exposure group. Since subjects are not randomized to exposure groups, it is important to control analytically for other risk factors that are associated with exposure status in the study (base) population. For example, we might want to control for the child's exposure to passive smoke at home; if other family members smoked during the mother's pregnancy, they are also likely to have smoked during the child's first 3 years of life. On the other hand, we should probably not control for birth weight even if it is a risk factor for the disease, because prenatal smoking affects birth weight. Thus, provided low birth weight is a risk factor for lower respiratory disease during the first three years of life, low birth weight is likely to be an intermediate variable in the causal pathway between prenatal exposure to passive smoke and the disease.

**Strengths of a Cohort Design.** The prospective cohort study is the observational design that is most similar to an experiment. The major strengths of this design derive from the fact that disease occurs and is detected after subjects are selected and after exposure status is measured. Thus, we can usually be confident that the exposure preceded the disease (i.e., there is no temporal ambiguity). This feature is particularly important when disease can also influence exposure status (e.g., persons with asthma moving to drier, less-polluted areas). Well-designed retrospective cohort studies also lack temporal ambiguity of cause and effect.

Another major strength of the cohort design is the usual lack of selection bias that threatens other basic designs (2). Disease status cannot, in principle, influence the selection of subjects except, perhaps, in poorly designed retrospective cohort studies. Sometimes researchers, ignoring this principle, propose random sampling to reduce bias. In fact, random sampling in a cohort study, unlike random

assignment, does not prevent or necessarily reduce epidemiologic bias in effect estimation; i.e., random sampling generally does not improve comparability between exposure groups. It does, however, make the study population representative of a larger, well-defined source population (sampling frame), which may make one's findings more generalizable. For example, suppose we initiated a prospective cohort study of lung cancer by mailing questionnaires to a random sample of 500,000 adults living in a given region served by population cancer registries. The questionnaire would request information on previous cancer diagnoses, exposure variables, and other risk factors for lung cancer. Following responses by 100,000 selected residents, the cancer registries would be used to identify all new cases of lung cancer diagnosed among respondents during the subsequent 5 years. Even though the 100,000 respondents will differ in many ways from the 400,000 nonrespondents, these differences will not cause epidemiologic bias in effect estimation. Nevertheless, the exposure effect observed for respondents (the base population) may not be generalizable to the population of nonrespondents. One possible reason for this lack of generalizability is that respondents and nonrespondents differ on the joint distribution of one or more effect modifiers.

As we will see in the next two sections, the same level of nonresponse in a cross-sectional or case-control study that we assumed in the above cohort example might seriously threaten the validity of effect estimation. Thus, unlike cohort (or randomized) studies, nonresponse in other basic designs can easily introduce selection bias because study cases have already occurred when subjects are selected. As noted in "Sources of Epidemiologic Bias," selection bias is most likely to be problematic when the investigator does not identify the base population from which study cases arose (as in cross-sectional studies and certain case-control studies).

**Weaknesses of a Cohort Design.** A potential weakness of cohort designs is the loss of subjects to follow-up due to death from other diseases, lack of participation, or migration. Unlike subject selection, loss to follow-up can easily bias effect estimation if attrition is associated with disease risk to a different extent for exposed and unexposed groups (2,35). Unfortunately, we can neither rule out nor confirm such bias by comparing lost subjects and followed subjects with respect to baseline characteristics (including risk factors) (35).

At best, baseline similarities between lost and followed subjects only suggest that loss to follow-up is probably not a major threat to validity, especially if the attrition rate is low.

Perhaps the major practical limitation of a cohort design, especially prospective studies, is its inefficiency for studying rare outcome events, which is what most diseases are in nonclinical populations. Because exposure status and other covariates must be observed at the start of follow-up in the entire study population, a rare disease would mean that most subjects will remain noncases. Comparing a small number of cases with a large number of noncases is statistically and economically inefficient because of the diminishing marginal return from additional noncases. Assuming a fixed sample size, therefore, it is more efficient to study a disease with an expected risk of 30% than to study a disease with an expected risk of 1%; the former will result in more precision and power for estimating and testing the exposure effect. Moreover, substantial increases in the sample size to compensate for too few expected cases is often impractical or impossible, especially when the size of the exposed population available for study is limited.

**Time-Dependent Exposures.** In conventional analyses of cohort-study data, exposure status and other covariates are usually treated as fixed variables measured at baseline. Yet the instantaneous and cumulative level of most environmental exposures changes during the follow-up period. Consequently, the greater the change and the longer the follow-up, the less appropriate are conventional methods of analysis. A common solution to this problem is to measure average exposure, duration of exposure, or cumulative exposure before and during the follow-up period; then these variables are analyzed like the simple baseline exposure variable, as possible (fixed) predictors of disease occurrence. Unfortunately, this approach also has methodologic problems: *a*) if the follow-up period for detecting disease overlaps the period during which exposure change is measured, the temporal relationship of an exposure-disease association is ambiguous. We may not know whether exposure changes preceded disease occurrence or disease preceded changes in exposure level. *b*) If the levels of exposure and/or other risk factors change over time, the associations between the exposure and these covariates also can change; then the amount of confounding of the estimated exposure effect will change. The analytic method described above, therefore, will not, in general, eliminate confounding due to these risk factors (even when there is no misclassification). *c*) When an extraneous risk factor affects subsequent exposure status and is

affected by previous exposure status, that risk factor can be a confounder and an intermediate variable simultaneously (36,37). For example, suppose we want to estimate the effect of exposure duration on mortality from a specific disease. If early symptoms of the disease lead to termination of exposure, then early symptoms, which is a risk factor for disease mortality, is both a confounder and an intermediate variable of the exposure-disease relationship. Consequently, standard methods of analysis will generally lead to a biased estimate of the exposure effect, whether or not one adjusts for the risk factor.

A statistical solution to the above problems was recently developed by Robins (36,37) who treats the prolonged or changing predictor variables as time-dependent covariates for which repeated observations are collected during the follow-up. The method involves estimating causal parameters for hypothetical exposure experiences of the study population (15). For example, we might want to compare the outcome risk for all subjects had they remained exposed throughout follow-up with these subjects had they remained unexposed, controlling for confounders at the start of each interval (time stratum).

### Cross-Sectional Study

A cross-sectional design involves a single ascertainment of disease prevalence in a study population that is usually sampled randomly from a single source population. In this sense, the source population is that larger group of individuals who are designated by the investigator as being eligible for inclusion in the study. Generally, in a cross-sectional study, we do not know how long prevalent (existing) cases have had the disease, nor can we identify the base population (at risk) from which the study cases arose. Exposure data on time-dependent variables are usually measured retrospectively to allow for expected variations in disease latency (before detection) and duration of expression (after detection).

The statistical analysis of cross-sectional data typically resembles the analysis of cohort or case-control data. Instead of comparing disease risks for exposed and unexposed groups, we compare disease prevalences ( $P$ ), as in a cohort study, or we compare the prevalence odds ( $P/(1-P)$ ), as in a case-control study (see "Case-Control Study"). Under certain conditions or assumptions, the prevalence ratio or prevalence odds ratio is approximately equal to the ratio of incidence rates or risks (i.e., the causal parameter of interest) (2,38). For example, disease prevalence in a population

is a function of both incidence and the duration of disease. If the mean duration of disease (from onset to recovery or death) is known to be identical for exposed and unexposed cases, we can be more confident that the prevalence odds ratio approximates the incidence rate ratio.

**Example.** Suppose we want to estimate the possible effect of prenatal exposure to passive smoke (as in "Cohort Study") on birth weight, categorized for convenience into low (<2500 g) and normal. We identify all live births delivered in one hospital during a given period (the source population); then we take a random or quasi-random sample (e.g., every third birth). By obtaining exposure data retrospectively from mothers near the time of delivery, we can compare the prevalence of low birth weight for infants prenatally exposed and unexposed to passive smoke, controlling analytically for confounders (e.g., maternal age, maternal smoking, and prenatal care).

Even though births may be regarded as incident events, the infant's weight at birth is a prevalence measure, because we do not know the size of the base population. The causal parameter of interest is a hypothetical comparison of retarded development between fetuses exposed to passive smoke and those fetuses had they not been exposed. Not only can we not observe this hypothetical condition of exposed fetuses being unexposed, but we do not (or cannot) follow the base population; the prevalence of low birth weight is simply the end result of that hypothetical follow-up.

**Strengths of a Cross-Sectional Design.** Because there is no follow-up, cross-sectional studies are less time-consuming and costly than prospective cohort studies. It is also feasible to examine many exposures and diseases in the same study, which makes this design useful for screening new hypotheses. In addition to causal inference, cross-sectional studies are important descriptively in health administration, planning, and policy analysis; information on disease prevalence is often required to assess the need and demand for health services and to evaluate intervention programs in specific target populations (2).

**Weaknesses of a Cross-Sectional Design.** A major methodologic limitation of many cross-sectional studies for making causal inferences is temporal ambiguity of cause and effect. Because we usually do not know the duration of the disease in prevalent cases and because exposure status is measured at the same time as disease status, often we cannot determine that exposure (or a certain accumulation of exposure) preceded disease occurrence. One approach for minimizing this problem is to collect retrospective expo-

sure data and information on previous medical diagnoses and the onset of symptoms associated with the disease under study. Not only may this approach be very uninformative for temporally linking exposure and disease, but it is also likely to worsen another potential problem, measurement error. Reliance on retrospective data increases the likelihood and magnitude of measurement errors, which generally leads to information bias. Furthermore, because all data are collected after disease has occurred, it is very possible for the error in measuring one variable to be related to the other variable (differential misclassification) or to error in measuring the other variable (correlated errors). Such possibilities are particularly likely in survey research and make potential information bias severe and unpredictable.

When cross-sectional studies are conducted without random sampling, they offer little opportunity for making statistical inferences about descriptive, population-specific parameters, e.g., the prevalence of a disease in a specified source population (28). The lack of random sampling may also worsen the potential problem of selection bias in effect estimation, which would be difficult to rule out a priori or to correct in the analysis. Even with random sampling, however, disease status or exposure status can influence the selection of subjects differentially by category of the other variable. For example, exposed cases may be less likely than others to be selected for study, perhaps because new exposed cases are less likely to survive than new unexposed cases (i.e., selective survival) or because exposed cases are less likely to enter the specified source population such as a hospital (i.e., Berkson's bias) (2). Similarly, selection bias can result from the differential participation of selected subjects (i.e., response bias).

### Case-Control Study

Case-control studies are distinguished from other basic designs by their sampling strategy: The investigator selects only a fraction of noncases (controls) from the population from which the cases were identified (2,3,5,34,39). Sometimes this population is not the true (primary) base population (out of necessity or convenience), and occasionally controls are assembled without regard for the identification of cases. The design may be longitudinal, involving incident cases, or cross-sectional, involving prevalent cases. In both types, the investigator establishes the ratio of controls to cases, which does not depend directly on the frequency of disease in the population. As in cross-

tional studies, exposure data on time-dependent variables are generally measured retrospectively to account for expected variations in disease latency.

**Estimation of Effect.** Unless the crude disease rate or the size of the base population is known, we cannot estimate the risk or rate of the disease in the exposed and unexposed populations. Nevertheless, we can estimate the effect of the exposure on disease by calculating the exposure odds ratio, which computationally is similar to the prevalence odds ratio in a cross-sectional study (2,3,19,20). For this estimation of effect to be valid, however, the controls must be representative of the base population that gave rise to the study cases. In this context, representative means having a similar distribution on other disease risk factors and indicators of disease detection. The best method for making the controls representative in this way is to sample them randomly (with or without matching) from the base population (see below).

**Matching.** As in any observational study, the investigator should control analytically for confounders by stratification or model fitting. Intuitively, it would appear that one method for achieving this control is to match controls to cases on extraneous risk factors (i.e., making controls similar to cases on the joint distribution of these risk factors). In a case-control study, however, it is not the matching alone that controls for the confounding effects of the matching variables; rather, stratification in the analysis eliminates this bias (1-6). In fact, the net effect of matching in case-control studies (but not in cohort studies) is to introduce selection bias that must be controlled in the analysis. Thus, if the matching is ignored in the analysis, the effect estimate will usually be biased (2,4,14).

The potential advantage of matching in the selection of subjects is that it allows the investigator to control for confounders more efficiently than if matching is not used (1-6). Yet, in this regard, matching can be counterproductive if one matches in a case-control study on strong correlates of exposure in the base population that are not risk factors (or proxy risk factors) for the disease. This type of overmatching results in a decrease in statistical efficiency (i.e., less precision for a given number of cases and controls) (1-6). The conditions for overmatching, however, are very different in cohort studies in which unexposed subjects are matched to exposed subjects (40). Matching can also be economically counterproductive for achieving a certain minimal precision if it costs more to match



than to increase the sample size without matching (41).

**Population-Based Case-Control Study.** In a population-based or hybrid case-control study, controls (noncases) are sampled directly from the base population that gave rise to the cases (39,42). When this design involves the follow-up of a large dynamic population, such as residents of a state, identification of new cases is usually based on data collected through a population registry. The validity of effect estimation depends on the completeness and accuracy of case ascertainment and on careful description of the base population. When the design involves the follow-up of individuals in a fixed cohort (e.g., as a part of a clinical trial or cohort study), identification of new cases is done by exams, interviews, or questionnaires administered periodically to each individual in the cohort during the follow-up. This latter strategy is now called a nested case-control study but also has been called a synthetic case-control study (43).

There are three alternative methods for selecting controls in a longitudinal, population-based case-control study: *a*) In density sampling, controls are selected longitudinally throughout the follow-up. Typically, they are individually matched to cases on time of each case's diagnosis or identification and possibly other factors; i.e., each control is known to be at risk (disease-free) at the time its matched case was first identified as diseased. An advantage of time matching is that exposure status is measured at about the same time for all subjects in each matched set (19). *b*) In cumulative sampling, all controls are selected at the end of the follow-up period during which cases are identified. Both cumulative- and density-sampling methods can be used even when controls are not selected directly from the base population. *c*) In case-base or case-cohort sampling, all controls are selected from the fixed base population at the start of the follow-up (42,44,45). An advantage of this method is that one control group can be used to study multiple diseases, provided that prevalent cases of each disease are excluded from the analyses involving that disease. In both case-base and density sampling, it is possible for a selected control to subsequently develop the disease and become a case in the study.

**Example.** Suppose we want to estimate the possible effect of prenatal exposure to passive smoke (as in previous examples) on the risk of sudden infant death syndrome (SIDS). Using hospital records and birth certificate information, we identify a large number of live births occurring in a given region during a certain period. Then this

base population is followed prospectively, using hospital records and/or a population registry to identify all infant deaths. For each diagnosed and confirmed case of SIDS, we randomly select two live controls matched to the case on age, race, and date of the case's death; thus, controls are density sampled from the follow-up experience (risk set) of the base population of live births. As soon as possible after case detection, we interview the mothers of all subjects to collect data on prenatal exposure to passive smoke and other covariates.

**Proportional (Case-Control) Study.** A proportional study is a special type of case-control study in which selected controls have developed or died from diseases other than the index disease under study (2). By definition, therefore, this is not a population-based design, since controls (especially deaths) may not be representative of the base population from which study cases arose. In a proportional morbidity study, both cases and controls are selected from a clinical population such as a hospital, clinic, physician's practice, or screening program. Controls are selected because they have other conditions or symptoms; thus, and they are likely to differ from the base population of cases in ways that affect disease occurrence or detection. This situation will usually occur when the exposure is a risk factor for those comparison diseases making up the control group. For example, we would obtain a severely biased estimate of the smoking effect in a hospital-based, case-control study if controls were selected from emphysema patients because smoking is a strong risk factor for emphysema.

Deaths comprise the entire population of a proportional mortality ratio (PMR) study. A group of deaths from the index disease (cases) is compared with a group of deaths from other diseases that might include selected comparison disease(s) or all other causes of death. Typically, all study deaths are identified retrospectively from the follow-up of a single population, such as persons living in a certain region or employed by a certain company during a given period. Although study deaths are incident events often identified from a defined base population, the outcome variable in this design is prevalence of disease at death; we do not have the proper denominator to estimate the disease-specific mortality rate in any (base) population. Furthermore, exposure data are not obtained for the base population but for study deaths only.

In the conventional proportional mortality study, comparison deaths are all other causes of death occurring in the population.

The traditional method of analysis is to compute the PMR, which is the proportion of exposed deaths resulting from the index disease divided by the proportion of unexposed deaths resulting from the index disease (6). Alternatively, the data are analyzed as in a case-control study; the researcher computes the mortality odds ratio (46,47). An important advantage of the alternative approach is that the comparison (control) group might be selected to include only those diseases thought to be unrelated to exposure status. This design strategy, which also should be used in a proportional morbidity study, can help reduce selection bias by making the comparison group more representative of the base population. Another advantage is that it allows use of the many analytic techniques developed for case-control studies (48,49).

#### **Strengths of a Case-Control Design.**

The major advantage of the case-control design over other basic designs is its efficiency for studying rare diseases, especially diseases with long latent periods. A greater proportion of study costs for collecting exposure and covariate data can be devoted to cases rather than expending most available resources on noncases. Thus, given a fixed sample size, case-control sampling in a study of a rare disease enhances the precision and power for estimating and testing the exposure effect. In addition, some case-control studies, particularly proportional mortality designs, tend to be relatively inexpensive and feasible because they can be based on readily available data sources.

#### **Weaknesses of a Case-Control Design.**

A key issue in the design of case-control studies is the method and procedures for selecting controls. Ideally, we would like to make each study population-based, such that every new case occurrence in a well-defined base population is immediately identified by the investigators and controls are sampled randomly from the base population. In practice, however, this goal is not so easily accomplished, especially when the base is a large, dynamic population that cannot be examined periodically. Even population surveillance and registry systems, when they exist, are likely to be very incomplete for many diseases, such as prostate cancer, Alzheimer's disease, and ischemic heart disease. If exposed cases are more likely or less likely to be detected or reported than unexposed cases, the resulting effect estimate will be biased. In a cohort study, this detection problem would manifest as differential disease misclassification bias; but in a case-control study, the detection problem produces a form of selection bias that might involve no disease

misclassification in the total sample and, therefore, cannot be corrected after subject selection (50). To prevent such detection bias, the investigator might select controls who, purportedly, underwent the same degree of medical surveillance as did study cases (51) (e.g., persons screened for the disease or patients treated for other related conditions). Unfortunately, this approach could introduce another problem by selecting for the control group individuals with other exposure-related conditions (see the discussion of proportional morbidity studies). The end result might be, for example, to overcompensate for potential detection bias, producing net bias in the opposite direction. In general, in the absence of perfect population-based methods, investigators must select controls to reflect the expected magnitudes of various potential selection problems.

When there is relatively little variability of exposure in the base population, we expect imprecise estimation of the exposure effect, even if the exposure is a strong risk factor for the disease. Although such inefficiency is usually quite apparent in cohort studies, it may not be so apparent in case-control studies, especially when the investigator does not know the exposure distribution in the base population. For example, if environmental exposure levels are high throughout the region of the base population, a comparison of cases and controls would result in an unstable estimate of effect and low power. As in cohort studies, the problem is not one of bias. Limited variability of exposure is likely to occur when exposure status for individual subjects is measured ecologically by assigning to each subject the exposure level observed for the area in which that subject lives or works. Other problems accompanying ecologic measurement are discussed in "Ecologic Designs."

**Two-Stage Designs.** Just as cohort studies are inefficient for studying rare diseases, case-control studies are inefficient for studying rare exposures. When both disease and exposure are rare, therefore, any basic design might require a very large sample size to ensure adequate power. One solution to this problem is a two-stage design: stage 1 is a basic design in which data are collected on exposure and disease variables only; in stage 2, covariate data and possibly more refined exposure data (with less measurement error) are collected on separate random samples of exposed cases, exposed noncases, unexposed cases, and unexposed noncases, all of which are identified from stage 1 results (52,53). Sampling fractions for stage 2 are set larger

for those exposure-disease groups that contain fewer subjects in stage 1. Thus, the investigator can obtain approximately equal numbers of the four exposure-disease groups in stage 2. Stage 1 results are used to estimate the crude (unadjusted) exposure effect, and stage 2 results are used to estimate the effect adjusted for covariates and possibly a more refined exposure effect. The analysis of stage 2 data considers the sampling fractions (52-56). The two-stage design is also advantageous when the cost of obtaining covariate data is large relative to the cost of obtaining exposure and disease data or when covariate data are missing on a majority of subjects (52,54).

**Case-Crossover Design.** A standard crossover design is an experiment or quasi-experiment in which each subject receives both the experimental and control treatments at different times (i.e., each subject serves as his or her own control) (57). Such designs are seldom used in environmental epidemiology because manipulation of treatment status (with or without randomization) is usually unethical or infeasible and because the outcome is usually a rare event. Recently, Maclure (58) proposed an observational analogue of the crossover study called the case-crossover design, which may be regarded as a special type of pairwise-matched, case-control study. This type of design can be used to estimate the possible transient effect of a brief exposure (e.g., coffee drinking) on the subsequent occurrence of a rare acute-onset disease (e.g., myocardial infarction) that is hypothesized to occur within a short time after exposure (i.e., during the effect period). All subjects are newly detected cases that serve as their own controls. That is, for each case, the observed odds of being exposed during the effect period (e.g., one hour before disease onset) is compared with the expected odds of being exposed during any random period of the same duration (assuming no exposure effect). The expected exposure odds is estimated from the subject's report of his usual exposure frequency before disease occurrence. For example, if a person drinks coffee twice each day and the effect period is 1 hr, the expected odds of exposure during any 1-hr period is 1/11. Thus, we would expect that, for every 12 cases who drank coffee twice each day, one case would have occurred by chance within 1 hr of exposure. Maclure recommends using standard methods of matched analysis for person-time (cohort) data to combine data from all cases; however, this approach needs further development to handle the temporal

autocorrelation of outcome status (i.e., a case is either exposed during the effect period or unexposed, but it cannot be both). Although the case-crossover design has not yet been used to examine the possible short-term effect of an environmental exposure, this type of study is feasible if we can measure such exposures.

## Genetic Study

The study of variation among individuals or groups in their sensitivity to environmental agents is one of the aims of environmental epidemiology. Such variation might be due to differences in host characteristics, including genetic factors, or to interactions with other exposures. A complete survey of the methods used to study the genetic determinants of disease would be beyond the scope of this report; instead, we will focus on the approaches that might be used to address the issue of gene-environment interactions.

Three basic types of information might shed light on the genetic component of such interactions: a classification of the subjects' genotypes at a major locus for disease susceptibility; some observable host characteristic (phenotype) that is genetically determined and linked with the genotype that was responsible for sensitivity; or family history as a surrogate for genetic (or shared environmental) influences. The choice of study design will depend upon which of these is sought.

The first is the most powerful approach, and its feasibility will grow as more and more genes are identified and assays for them become available. If the genotype is observable, it can be considered simply another risk factor and any of the basic design and analysis strategies used in epidemiology are applicable. For example, Caporaso et al. (59) reported a case-control study of lung cancer, in which the rate of metabolism of the antihypertensive drug debrisoquine was taken as a phenotypic marker for a gene in the cytochrome P450 system that is responsible for metabolism of carcinogens. It was shown that intermediate and high metabolizers were at higher risk of lung cancer overall and that there was an interaction between metabolic rate and exposure to occupational carcinogens and smoking. In this example, the genotype was not observed directly but inferred from the phenotype; but recent advances in molecular genetics, such as the use of restriction fragment length polymorphism, are making direct observation increasingly feasible.

Identifying host characteristics that interact with environmental exposures can be done in essentially the same way. A



familiar example might be skin color as a marker for sensitivity to sunlight in the production of melanoma. No extensions of standard epidemiologic methods would be needed to address this question. The only subtlety in this case arises when the gene determining the host characteristic is not the disease susceptibility locus but only linked to it (i.e., nearby on the same chromosome). A particular marker allele might be associated with the disease in one family and a different allele in another family; but in both families, the marker and the disease would be inherited together. This possibility requires family data and the techniques of linkage analysis. To date, such analyses have been applied only to the study of genetic effects without reference to the environmental covariates, but statistical techniques that would assess such combined effects recently have become available (60).

Before trying to identify a specific major gene that is related to sensitivity to environmental exposures, one should assess whether there is any evidence that such sensitivity has a genetic basis. This also requires collection of family data, but unlike the standard analyses aimed at examining the main effect of genetics, one would also want to examine interactions between family history and environmental exposures. Geneticists commonly assemble a small number of very large pedigrees, sometimes selected to maximize the chances that a major gene is operating in the families, and subject them to segregation analysis to study the mode of inheritance. Again, these analyses seldom account for environmental covariates and interactions, although such methods are now available. In contrast, epidemiologists begin with a large population-based series of cases and controls and restrict attention to the first- and sometimes second-degree family members. Their analyses usually are limited to a simple family history covariate (e.g., presence of an affected member, number of affected members, etc.) in standard multivariate risk-factor models, possibly but seldom including interactions with environmental covariates. Susser and Susser (61) have discussed two basic approaches to such data: in the case-control approach, cases are compared with controls in terms of their family histories; in the cohort approach, the incidence of disease in the exposed (case) families is compared with the incidence in unexposed (control) families. Either approach could easily be extended to incorporate environmental covariates and their interactions with family history. The only difference is that the cohort approach requires covariate data on all of the family members, whereas the

case-control approach requires data only on the originally sampled cases and controls. Clearly, the cohort approach is more informative, but the conditions under which the additional effort warrants the gain in information about gene-environment interactions have not been investigated.

The two major design issues to be addressed in such studies are the method of ascertainment of families and the information to be collected on family members. The former has been discussed at great length in the genetics literature. The basic problem is that if families are ascertained through affected probands, families with multiple cases will tend to be overrepresented. Therefore, various corrections for ascertainment are applied in the standard methods of genetic analysis. The relevance of these approaches to the epidemiologic designs for gene-environment interactions requires further research. Often, only very limited information is collected about family history in epidemiologic studies. The minimal information should be an enumeration of all affected family members together with the sex and age of each family member at risk; as discussed above, information on major risk factors for all family members at risk may also be desirable. Because larger and older families are likely to have more familial cases, the presence or number of familial cases is not suitable as a family history covariate. Moreover, expressing the number affected as a proportion does not solve the problem because multiple cases in large families are more informative than single cases in small families. A more appropriate comparison is between the observed number of familial cases and the expected number based on the person-time at risk, in which the comparison is adjusted for age, sex, and other important risk factors (62).

## Ecologic Designs

An ecologic or aggregate study is one in which exposure levels of individuals are not linked to disease occurrence of those individuals. The net result is that the unit of statistical analysis is usually the group, typically persons living in a geographic area such as a census tract, county, or state. For each group or region, therefore, we know the average exposure level or distribution and the disease rate, but we do not know the joint distribution of these two variables. Given a dichotomous exposure, for example, we would not know the numbers of exposed and unexposed cases in each group. Thus, we cannot estimate the exposure effect directly by comparing the disease rate for exposed and unexposed populations.

Ecologic designs are therefore incomplete (2) in the sense that they lack certain information ordinarily contained in the basic designs. As noted in "Problems in Environmental Epidemiology," the primary reason for this missing information in environmental epidemiology is our inability or lack of resources to accurately measure environmental exposures in large numbers of individuals. Thus, the widespread use of ecologic designs in environmental epidemiology reflects a fundamental problem of exposure measurement. In addition, ecologic studies represent an inexpensive design option for linking available data sets or record systems, even when exposures are measured at the individual level. The appeal of this alternative is that aggregate summaries of many exposures, including sociodemographic and other census variables, are often available for the same regions that are used to summarize morbidity and mortality data.

With the inclusion of covariate data in an ecologic study, the analysis may be only partly ecologic. This condition occurs when the joint distribution of two or more, but not all, variables is known within groups. For example, suppose we want to examine the possible effect of radon exposure on lung cancer incidence, controlling for age (the covariate). Although we might know the age distribution of all new cases and all persons at risk within each county (from tumor registry and census data), we would usually not know the within-county association between radon exposure and the other two variables. Sometimes data sets like this are analyzed with the individual as the unit of analysis, where each individual is assigned the average radon exposure level that was measured for the region in which he or she lives. Such ecologic measurement of exposure means that there is likely to be substantial error in measuring the individual's exposure to radon, which could result in information bias of effect estimation.

## Types of Ecologic Studies

Ecologic studies may be classified into five design types that differ in several ways, including methods of subject selection and methods of analysis (2,63).

**Exploratory Studies.** In exploratory ecologic studies, we compare the rate of disease among many contiguous regions during the same period, or we compare the rate over time in one region. In neither approach are exposures to specific environmental factors measured (for individuals or groups). The purpose is to search for spatial or temporal patterns that might suggest

an environmental etiology or more specific etiologic hypotheses.

The simplest type of exploratory study of spatial patterns is a graphical comparison of relative rates across all regions (i.e., mapping study), possibly accompanied by a statistical test for the null hypothesis of no geographic clustering (64). In mapping studies, however, a simple comparison of estimated rates across regions is often complicated by two statistical issues. First, regions with smaller numbers of observed cases show greater variability in the estimated rate; thus, the most extreme rates tend to be estimated for those regions with the fewest cases. Second, nearby regions tend to have more similar rates than do distant regions (i.e., positive autocorrelation). A statistical method for dealing with both complications involves empirical Bayes' estimation of rates using an autoregressive spatial model (65).

In certain exploratory studies of spatial patterns, regions are characterized in terms of general ecologic indicators such as degree of urbanization (urban versus rural), degree of industrialization (agricultural versus nonagricultural), population density, socioeconomic status, and ethnic diversity. The analysis of these data usually involves comparisons of regions grouped by one or more ecologic indicators. This approach resembles the statistical methods used in multiple-group studies (see "Interpretation of Results").

An exploratory ecologic study was conducted by Mahoney et al. (66), who compared age-standardized mortality ratios for cancers, by sex, among all cities and towns in New York State (exclusive of New York City) between 1978 and 1982. By grouping these regions by quintile of population density, they examined the associations between density and deaths from all cancer sites and selected sites, by sex. They found linear associations between increasing population density and total cancer mortality in both men and women. Because population density may reflect various risk factors for different cancers, the authors acknowledge that their findings are consistent with several alternative explanations.

An exploratory study of temporal patterns is generally done by comparing disease rates for a geographically defined population over a period of at least 20 years. A common statistical or graphical approach for analyzing such longitudinal data is cohort analysis (not to be confused with the analysis of data from a cohort study) (2). The objective of this approach is to estimate the separate effects of three time-related variables on disease occurrence: age, period (calendar time), and

birth cohort (year of birth). Because of the linear dependency of these three variables, there is an inherent statistical limitation (identification problem) with the interpretation of cohort-analysis results. The problem is that each data set has alternative explanations with respect to the combination of age, period, and cohort effects. The only way to decide which interpretation should be accepted is to consider the findings in light of other (prior) knowledge of the disease and its determinants.

A cohort analysis was conducted by Lee et al. (67) on melanoma mortality among white males living in the United States between 1951 and 1975. They concluded that the apparent increase in the melanoma mortality rate during that period was due primarily to a cohort effect. That is, persons born in more recent years carried with them throughout their lives a higher mortality rate than did persons born earlier. In a subsequent review paper, Lee (68) speculates that the cohort effect might reflect the impact of changes in a major risk factor operating during youth, such as sunlight exposure or burning.

**Space-Time Cluster Study.** Space-time clustering refers to the interaction between place and time of disease occurrence, such that cases that occur close in space also occur close in time (2). Evidence of space-time clustering may suggest person-to-person transmission of an infectious agent or the effects of point-source exposures, depending on the disease and the cluster pattern. The analytic search for space-time clusters requires special statistical techniques that may or may not incorporate information on the base population and covariates (69,70). Although the unit of analysis for these methods is usually the individual, space-time cluster studies are classified here with ecologic designs because closeness in space and time is a proxy measure for environmental exposures—or at least the opportunity for exposures. Thus, use of place and time information is analogous to use of spatial or temporal indicators in the exploratory study.

Space-time cluster analyses may be used when members of a community perceive a cluster or excess number of cases of one or more diseases in their area. This activity is often motivated by the suspicion that the apparent cluster is caused by a specific environmental exposure, such as chemical waste, pesticides, or electromagnetic fields. When investigation begins, the first steps are to verify the diagnoses of all reported cases and identify any additional cases in the cluster area, which must be defined. In addition to space-time cluster analyses, the investigators

will probably want to compare the disease rate in the cluster-area population with the rate in another population thought to be unexposed (retrospective cohort study), and they may conduct a population-based case-control study to identify risk factors for the disease.

**Multiple-Group Study.** In a multiple-group ecologic study, we assess the ecologic association between average exposure level or prevalence and the rate of disease among many groups or regions. This is the most frequently used ecologic design in environmental epidemiology. Studies are usually conducted by linking separate sources of data. For example, census and tumor-registry data might be combined to estimate cancer rates for all counties in a state; other state records or surveys might be used to estimate average exposure levels by county. Statistical methods for estimating exposure effects in multiple-group studies are discussed in "Interpretation of Results" and by Prentice and Thomas in this issue.

Hatch and Susser (71) conducted a multiple-group ecologic study to examine the association between background gamma radiation and childhood cancers between 1975 and 1985 in the region surrounding the Three Mile Island nuclear plant. Using data from a 1976 aerial survey, they estimated the average radiation level for each of 69 tracts in the study region. The results of their analyses showed a positive association between radiation level and the incidence of childhood cancers. The authors were cautious in making causal inferences, however, because the large effect observed for solid tumors, as well as leukemias, was not expected.

**Time-Trend Study.** In time-trend (or time-series) studies, we assess the ecologic association between change in average exposure level or prevalence and change in disease rate in one geographically defined population. The assessment may be done by simple graphical displays or by more formal statistical techniques (72–75). With either approach, however, the interpretation of findings is often complicated by two issues. First, changes in disease classification and diagnostic criteria can produce very misleading results. Second, the latency of the disease with respect to the exposure of interest may be long, variable across cases, and/or unknown to the investigator; thus, employing an arbitrary or empirically defined lag between the two trends can also produce very misleading results (76).

Darby and Doll (77) compared the trends of average annual absorbed doses of radiation fallout from weapons testing and childhood leukemia rates in three European countries between 1945 and 1985. Although the

leukemia rates varied over time in each country, they found no convincing evidence that these changes were attributed to changes in fallout radiation.

**Mixed Study.** The mixed ecologic design combines the basic features of the multiple-group study and the time-trend study. The objective is to assess the ecologic association between change in average exposure level or prevalence and change in disease rate among many groups. Thus, two types of comparisons are made simultaneously: change over time within groups and differences among groups.

For example, Crawford et al. (78) evaluated the hypothesis that hard drinking water (i.e., water containing more calcium and magnesium ions) is a protective risk factor for cardiovascular disease (CVD). They compared the absolute change in CVD mortality rate between 1948 and 1964, by age and sex, in 83 British towns. The towns were divided into three groups: *a*) five had experienced increases in water hardness; *b*) six had experienced decreases, and *c*) 72 had experienced little or no change in water hardness. In all sex-age groups, especially for men, the authors found an inverse association between trends in water hardness and CVD mortality. In middle-aged men, for example, the increase in CVD mortality was less in towns that made their water harder than in towns that made their water softer.

### Interpretation of Results

Statistical analysis in a multiple-group study usually involves fitting the data to a mathematical model (see Prentice and Thomas, this issue). The outcome variable is a function of the disease rate in each group; predictors include the average exposure level or proportion exposed in each group plus other ecologic covariates, the effects of which the investigator wants to control. We show in "Control for Covariates" that these covariates need not be confounders (i.e., at the individual level within groups).

Results of the fitted model can be used to estimate the exposure effect, i.e., the same causal parameter we would like to have estimated had the study been conducted at the individual level (63,79,80). For example, suppose the exposure variable is the proportion exposed in each group and there are no covariates. Assuming a linear model, we can use weighted least-squares regression to estimate the slope (*b*) and intercept (*a*). The predicted disease rate in a group that is entirely exposed is then  $a + b(1) = a + b$ , and the predicted rate in a group that is entirely unexposed is

$a + b(0) = a$ ; therefore, the estimated rate ratio is  $(a + b)/a = 1 + b/a$ . It is important to note that this estimation procedure implies extrapolating the results of the model to both extreme values of the exposure variable, either or both of which may lie well beyond the observed range. It is not surprising, therefore, that different model forms can lead to very different estimates of effect (81). In fact, certain model assumptions may lead to rate-ratio estimates that are negative and thus meaningless.

### Ecologic Bias

The use of ecologic data to estimate causal parameters has a major methodologic limitation, called the ecological fallacy (82), aggregation bias (83), cross-level bias (84), and ecologic bias (85,86). Ecologic bias refers, in general, to the failure of ecologic estimates of effect to reflect the true effect at the individual level. Some of this bias may occur in individual-level studies of the same population, but some of it is due specifically to the aggregation of subjects into groups. More importantly, the magnitude of ecologic bias is likely to be more severe and less predictable than is individual-level bias in estimating the same effect (63,81,86,87). It is very possible, for example, that an ecologic analysis of a (true) positive risk factor would produce an apparently protective effect.

The underlying problem of ecologic bias may be regarded as a special form of information bias resulting from within-group heterogeneity of exposure status, which is not captured in the analysis. For example, a positive linear relationship between proportion exposed and disease rate does not necessarily mean that exposed individuals are at greater risk for the disease than are unexposed individuals; rather, unexposed individuals may be at greater risk in groups containing proportionally more exposed individuals. The implication of this latter explanation is that an individual's group affiliation has an effect on disease occurrence that reflects more than just the individual's exposure status.

A mathematical understanding of ecologic bias was first provided for correlation coefficients by Robinson (88) and later extended to regression coefficients by Duncan et al. (89). Nevertheless, the conditions for valid ecologic estimation and the relationship between ecologic bias and other methodologic issues are still not well understood. Because the results of ecologic analyses are often used to influence policy decisions, as well as to make causal inferences, it is important for researchers to appreciate the complexities of ecologic inference.

**Sources of Ecologic Bias.** Ecologic bias is often confused with confounding, perhaps because regional differences in disease rates can be due to variation in the distribution of extraneous risk factors across regions. To clarify the confusion between these two concepts, Greenland and Morgenstern (86,87,90) show that ecologic bias can arise from three different sources.

**Within-Group Confounding (At the Individual Level).** The exposure effect may be confounded within groups (as described for nonecologic studies in "Sources of Epidemiologic Bias"). Thus, if the within-group effect is equally confounded by the same unmeasured risk factors in every group, we can expect the ecologic estimate of effect to be biased as well. In general, ecologic estimates will be biased in this way if the net within-group bias across groups (due to uncontrolled confounders) is not zero. It is possible, therefore, for positive confounding in certain groups to cancel negative confounding in other groups.

The other two sources of ecologic bias are unique to this design and can be understood by considering group (or group affiliation) as a nominal predictor of disease occurrence at the individual level.

**Confounding by Group.** Ecologic bias can occur when the disease rate in the unexposed population varies across groups. Since average exposure level also typically varies across groups, group is a confounder of the exposure effect at the individual level. This set of conditions may occur if one or more unmeasured risk factors are differentially distributed across groups, even if these risk factors are unrelated to exposure status within groups and, therefore, are not confounders at the individual level.

**Effect Modification by Group.** Ecologic bias can also occur when the exposure effect varies across groups, i.e., when group modifies the effect of the exposure at the individual level. This condition may result from extraneous risk factors (effect modifiers) being differentially distributed across groups or by misspecification of the model form used to analyze the data. Ecologic bias of this type tends to be more severe when there is little variability in average exposure across groups (85), even when the effect modification is relatively weak and there is no confounding by group.

Taken together, the above principles imply that there will be no ecologic bias if the disease rate in the unexposed population and the exposure effect do not vary across groups and if there is no net confounding within groups. Unfortunately, it is very unlikely that all of these conditions

will be met in one ecologic study. Although small departures from these conditions may result in substantial bias (81,86), it is also possible that there will be little or no bias in certain studies when one or more of these conditions are not met.

If every group were completely exposed or unexposed, there would be no ecologic bias attributable to confounding or effect modification by group. Indeed, if all covariates were measured at the individual level, such a study would not be an ecologic design. Thus, to reduce ecologic bias, we should select regions that minimize within-region exposure variation and maximize between-region variation (63,81). One strategy for achieving these goals is to choose the smallest unit of analysis for which required data are available (e.g., census tracts or blocks). Unfortunately, certain data are seldom available at this level (e.g., personal behaviors and biomedical factors), and there is no guarantee that these smaller units are more homogeneous with respect to exposure status. Furthermore, use of smaller groups might increase the problem of migration between groups (see "Other Methodologic Problems").

### Control for Covariates

In a study conducted entirely at the individual level, an extraneous risk factor produces bias (confounding) in effect estimation only if it is associated with exposure status in the base population (see "Sources of Epidemiologic Bias"). In a multiple-group ecologic study, however, an extraneous risk factor can produce ecologic bias even if it is not associated with exposure status within regions (at the individual level) (86,87,90). Such bias occurs typically because the ecologic association (across regions) between the exposure and risk factor produces confounding and/or modification of the exposure effect by group (see "Ecologic Bias"). Conversely, an extraneous risk factor that is a confounder within regions may not produce ecologic bias if the net within-group bias is zero (see "Ecologic Bias") or if the risk factor is ecologically uncorrelated with the exposure.

One method to control for extraneous risk factors in ecologic studies is to include predictor terms for these risk factors in the model (e.g., the proportion of smokers or the mean family income in each region). Unfortunately, even when such covariate data are available for all regions, ecologic adjustment usually cannot be expected to remove completely the ecologic bias produced by these risk factors. In fact, it is possible for such ecologic adjustment to increase bias (86).

The general conditions under which the ecologic control for extraneous risk fac-

tors either increases or decreases bias have not been delineated. Yet, under certain restrictive conditions, ecologic control for covariates will produce unbiased estimates of the exposure effect, provided there are no other sources of bias (e.g., outcome misclassification). If the effects of the exposure and the covariate on disease rate are exactly additive within every region (i.e., the rate difference for each variable is constant across levels of the other variable) and if the rate conditional on both predictors is exactly the same in every region, ecologic regression of disease rate on the mean exposure and covariate levels (i.e., multiple linear regression) will lead to unbiased estimates of both effects (83,84). Under these conditions, group affiliation does not confound or modify the exposure effect at the individual level. However, as shown by Greenland (81), relatively minor deviations from perfect additivity (linearity) can lead to appreciable ecologic bias because ecologic rate ratios can be extremely sensitive to the choice of model form, in contrast to individual-level estimates. Furthermore, the two conditions noted above are only sufficient for no ecologic bias to occur; ecologic bias may be absent when either or both conditions are not met.

Richardson and Hémon (91) recently pointed out that there is another set of conditions for which ecologic control of covariates is possible. If *a*) the exposure and covariates are uncorrelated within regions, *b*) their effects on disease are multiplicative (i.e., the rate ratio for each variable is constant across levels of the other variable), and *c*) the rate conditional on both predictors is exactly the same in every region, then ecologic bias due to the covariates can be removed or largely reduced by including product terms in the linear model. Of course, such conditions are very difficult to verify in ecologic studies; if the exposure and covariates (other risk factors) are correlated within regions, the covariates will be confounders at the individual level and substantial ecologic bias can occur even with product terms in the model (81).

When the data are not entirely ecologic (see "Ecologic Designs"), rate standardization is another method often employed to adjust for extraneous risk factors in ecologic studies. For example, if the age distribution is known for cases and for the base population in every region, we can mutually standardize the rate in every region to the age distribution of a well-defined (standard) population (5); then we use the standardized rates as the outcome variable in the ecologic analysis. Unfortunately, this method does

not always reduce ecologic bias due to the variables for which the rates are standardized; in fact, the result may be to increase bias appreciably (86,92). Standardization can be expected to reduce ecologic bias only if all variables in the model (i.e., disease and all predictors) are mutually standardized for those other confounders (e.g., age) not included as predictors in the regression model. This method is often not feasible, for example, when the investigator does not know the age distribution of exposed and unexposed populations within every region.

### Other Methodologic Problems

In addition to ecologic bias and the related difficulties of controlling for extraneous risk factors, there are other methodologic problems with ecologic analysis, a few of which are addressed below.

**Exposure Misclassification Bias.** As noted in "Sources of Epidemiologic Bias," nondifferential misclassification of exposure status in individual-level studies nearly always results in bias toward the null value; e.g., the estimated rate ratio will be closer to one than is the true rate ratio. In multiple-group ecologic studies, however, this principle does not hold when the exposure variable is formed from the aggregated observations of all individuals in each region (e.g., the proportion exposed). Brenner et al. (93) have shown that nondifferential misclassification of a binary exposure within groups usually leads to overestimation of the rate ratio (away from the null value) in ecologic studies, which can be severe. This apparent contradiction between ecologic and individual-level studies can be understood by considering just two regions. Nondifferential exposure misclassification in both regions will produce an estimated difference in exposure prevalence that is smaller than the true difference. Consequently, the estimated regression coefficient (slope) for the exposure variable in a linear ecologic model will be overestimated, leading to overestimation of the rate ratio. Little is known about the impact in ecologic studies of within-group error in measuring continuous or multiple-category exposures.

**Confounder Misclassification.** In studies conducted at the individual level, misclassification of a confounder, if nondifferential with respect to exposure and disease, will usually reduce our ability to control for the confounder in the analysis (94,95). That is, adjustment will not completely eliminate the bias due to the confounder. In ecologic studies, however, nondifferential misclassification of a binary confounder within groups does not affect our ability to control for that confounder (96). Thus, sur-

prisingly, nondifferential misclassification of a confounder is less problematic in ecologic studies, provided there is no ecologic bias, than in individual-level studies.

**Collinearity.** It is probably more common in ecologic studies than in other studies for two or more predictors to be highly correlated across groups (63,97,98). This issue is particularly relevant with environmental factors, such as the associations between levels of different contaminants in air or drinking water or associations between different socioeconomic indicators. The implication of such collinearities is that it is very difficult, perhaps impossible, to separate these effects statistically; analyses yield model coefficients with very large variances and often severely distorted estimates of effect.

**Temporal Ambiguity of Cause and Effect.** Use of incidence data in a cohort study usually implies that disease occurrence did not precede exposure to the hypothesized risk factor. Yet, in multiple-group or time-trend ecologic studies use of incidence data provides no such assurance against this temporal ambiguity (63). This inferential problem is most troublesome when it is possible for disease to influence exposure status either at the individual level (see "Cohort Study") or at the ecologic level (e.g., interventions designed to reduce exposure levels in areas with high rates of disease).

The problem of temporal ambiguity in ecologic studies is further complicated by an unknown or variable latent period between exposure and disease occurrence. The investigator can only attempt to deal with this problem by establishing a specific lag period between observations of average exposure and disease rate. Even when the average latency is known, however, appropriate data may not be available to accommodate the desired lag.

**Migration.** Migration of individuals into or out of the base population can cause selection bias in any type of epidemiologic study, because migrants and nonmigrants may differ on both exposure prevalence and disease risk. Little is known about the magnitude of this bias or how it can be reduced in ecologic studies, especially when studying diseases with long latent periods. One approach might be to use larger geographic groups (e.g., states instead of counties as units of analysis) (99). Unfortunately, this approach is also likely to increase the potential for severe ecologic bias, because it makes the groups less homogeneous with respect to exposure (see "Ecologic Bias"). Another approach might be to incorporate available data on the distributions of residential durations within regions, but this

approach needs more work to provide a reliable method of bias reduction.

## Current Issues and Recommendations

A general goal of epidemiologic research is to obtain the most information about possible health effects with minimal and/or available resources. Given the difficulties in estimating effects of specific environmental exposures in human populations, this goal is not easily obtained and optimal research strategies are not readily identified. Below, we highlight several current methodologic issues in environmental epidemiology and make some recommendations for future work.

**Study Design.** No single design best meets the objectives of every epidemiologic study. In practice, study objectives are shaped by many factors—current knowledge, previous findings, institutional mandates, societal values, personal preferences, etc. Although a prospective cohort study might be expected, in general, to produce less bias than would a hospital-based (proportional) case-control study, the latter design might be a rational choice in certain situations. Even an ecologic design, despite its limitations, might be appropriate; it may be the only practical option at a given time.

The challenges of environmental epidemiology, therefore, cannot be solved simply by advocating the use of certain, more expensive study designs. In addition to committing more resources to the conduct of epidemiologic research, we need to develop new designs to meet specific objectives more efficiently. For example, in "Case-Control Study," we discussed the use of two-stage designs to investigate associations between rare diseases and rare exposures and to control for covariates that are relatively expensive to measure. New approaches are also needed to identify intermediate variables in observational designs, to evaluate interaction effects (effect modification) more efficiently, and to deal with the problems of nonparticipation, nonresponse, and noncompliance. Another need in environmental epidemiology is to understand better the relationship between acute biological changes and chronic health effects. For example, we might combine experimental and observational methods to determine the extent to which short-term changes in pulmonary function caused by exposure to air pollutants lead to chronic respiratory disease (2).

**Bias Reduction.** In nonrandomized studies, it is important for the investigator to deal with confounding in the analysis. This is achieved by identifying potential

confounders in the design phase and measuring them accurately in the study population. The prevention of selection bias, however, is not so straightforward because it depends on identifying all cases that occur in a well-defined (base) population at risk. When new cases occur infrequently or when it is otherwise impractical to re-examine enough individuals to detect all new events, the prevention of selection bias depends on population surveillance and monitoring systems, such as population-based tumor registries and industrial surveillance. Although these systems may be expensive to implement and operate, they are often necessary to reduce the threat of selection bias.

Unfortunately, population-based systems may not be sufficient to prevent selection bias with diseases for which detection depends critically on care seeking, symptom reporting, and complex differential diagnoses. A key problem is that not all persons with an illness recognize their symptoms and seek medical attention. Thus, exposure effects observed for these diseases in epidemiologic studies might reflect the effects of the exposure on illness behavior as well as the effects on illness occurrence (100).

Another solution to incomplete or inadequate case detection is to control analytically for methodologic covariates that reflect differences in illness behavior. For example, we might measure the individual's tendency to seek medical care and treat this variable as a confounder. The measurement of this covariate should be independent of disease status; otherwise, covariate adjustment will probably lead to bias toward the null value. This approach needs further development and evaluation.

An alternative strategy for studying diseases that are difficult to detect in large populations, such as musculoskeletal conditions, is another type of two-stage design. In the first stage, a large population is surveyed cross-sectionally or longitudinally by questionnaires or interviews to identify persons with symptoms characteristic of the disease. The second stage involves case-control sampling of the population to compare persons with and without these symptoms (i.e., cases and controls). In this stage, subjects are given more definitive diagnostic tests to identify true cases of the disease. By comparing diagnostic test results between selected cases and controls, the investigators can assess the validity of their symptom-based criteria, suggest improvements in clinical diagnosis, and estimate exposure effects. The latter objective requires the development of statistical methods appropriate for the sampling strategy.

**Quality of Measurement.** As noted earlier (see "Problems in Environmental Epidemiology"), a major challenge in environmental epidemiology is to measure accurately each individual's exposure to suspected and known risk factors for the disease under study. In the absence of previously validated and inexpensive methods for measuring exposures and covariates in large groups, it has become common practice to use more than one method or source of information to measure these variables. Nevertheless, it usually is not clear how different methods or sources of information should be combined or what data should be combined to minimize measurement errors and estimation bias (4,101,102). We need more methodologic research in this area to provide guidelines for the measurement of specific exposures in particular types of populations. One approach that might be pursued with environmental exposures is to combine ecologic data with self-reported data on individual behaviors. For example, suppose we collect ecologic data on pesticide spraying and distribution throughout a large region. We could then obtain from subjects the location of their homes; the type and location of their work; their use of drinking water; and how often they swim, fish, and participate in other activities that would affect their exposure to pesticides in the region.

Frequently, an accurate method does exist for measuring an exposure, but the application of this method to all subjects in a population is prohibitively expensive or infeasible. In such cases, many investigators rely on less accurate methods for the total sample and use the more accurate method in a subsample of subjects. Assuming the accurate method is perfectly valid (i.e., the gold standard), the results of the validation sub-study are used to quantify the amount of measurement error, which is then used in the total sample to correct for misclassification bias involving the imperfect measure of exposure. Some important issues need to be considered to make this approach advantageous. First, how many subjects and what proportion of the total sample should be included in the substudy (103,104)? Second, how should we correct for exposure misclassification in the analysis, especially when the accurate method may not be perfectly valid or when the subsample is not representative of the total sample (see also Prentice and Thomas, this issue)?

**Ecologic Inference.** Because of inherent problems of measurement, most epidemiologic studies of environmental exposures are at least partly ecologic. When all data, except a single exposure, are obtained at the individual level, however, the ecologic

problem amounts to possible misclassification bias, which is well understood and often predictable. Yet, when the unit of analysis is the group, the resulting ecologic bias is far less predictable and can be relatively severe in magnitude, especially when other sources of bias are present. Thus, in general, ecologic analyses do not provide very accurate estimates of effect. To make ecologic findings more informative, therefore, we need more theoretical work to specify the conditions for which ecologic estimates can be expected to be reasonably valid. With this information, we might then collect additional data to check those key assumptions or to correct ecologic estimates. For example, by obtaining detailed individual-level data on the exposure and certain covariates in samples of selected groups, we might be able to determine the limits of ecologic bias in estimating the exposure effect (see "Control for Covariates" and Prentice and Thomas, this issue).

Essentially, ecologic bias (aside from within-group bias) occurs because group affiliation or the average exposure level of the group affects disease occurrence independently of exposure status at the individual level (see "Ecological Bias"). The structural effects of such ecologic variables, if they can be separated from other effects at the individual level, might be informative, rather than just a source of error. Thus, by including both ecologic and individual-level predictors (possibly of the same exposure) in the analysis, we might enhance our understanding of disease occurrence. This type of contextual or multi-level analysis has been used extensively in social science research (105–108) but rarely in epidemiologic research (109). In addition, if the effect of a risk factor is known from previous research, the results of an ecologic analysis involving that risk factor could be used to evaluate the potential or realized impact of a population intervention, which may not be completely estimable at the individual level (63). A more profound understanding of ecologic bias, therefore, could yield benefits to other public-health research.

**Gene–Environment Interactions.** Because both genetic and environmental factors contribute to the etiology of most diseases, we would typically expect factors of each type to confound and/or modify the effect of the other. We know, for example, that a combination of both environmental/personal factors and genetic susceptibilities are sufficient for the development of certain diseases. Yet standard methods of epidemiologic research and population genetics have not been well integrated (110). As indicated in "Ecological Bias," we need new methods for incorporating

environmental variables in genetic (e.g., linkage) analyses of pedigree data. We also need to understand better the relationship between those parameters estimated in pedigree studies and the effect parameters estimated in standard epidemiologic studies; and we need to understand better how the estimates of gene–environment interactions in pedigree studies are biased by confounding, measurement error, and family selection (ascertainment). With this understanding, we can devise new methods to prevent or control bias. Analogously, the use of family data in standard epidemiologic designs (e.g., history of disease and/or its risk factors in relatives) requires further development in order to handle differences in family size and composition among subjects. With the recent advances in molecular genetics, the integration of epidemiology and population genetics is likely to become more important in the future.


**Sample Size and Power.** As noted in "Problems in Environmental Epidemiology," epidemiologic studies of environmental exposures often require large sample sizes to detect risk-factor effects with sufficiently small statistical error. To address this concern, researchers are usually expected to justify their proposed sample size by estimating the power of their study for testing one or more major hypotheses (i.e., the probability of detecting an association of at least a certain magnitude with a designated Type I error—alpha level typically set at 5%). This is a rather straightforward procedure when the power estimation is applied to two dichotomous variables (exposure and disease) (1,4,111). Yet all observational studies require more complicated analyses to make causal inferences — e.g., to deal with polytomous, continuous, or time-dependent exposures; covariate adjustment; the assessment of interaction effects; matching; and other special design features. Although methods of power estimation do exist for many of these complicating features, they require additional specifications (assumptions) about which the investigator is not likely to have adequate information. Further development of these methods would be useful, therefore, to identify techniques that are both practical and informative in specific situations, including ecologic studies for which sample size requirements have received little attention.

One parameter the investigator must specify to justify the proposed sample size is the magnitude of effect expected in the data or the minimum effect regarded as important to detect. In the absence of previous epidemiologic studies involving similar exposure levels, the expected effect is generally specified rather arbitrarily (e.g., a



rate ratio of 2). Sometimes, however, there are exposure-response findings from animal studies or occupational studies with higher exposure levels, which could be used to estimate the environmental exposure effect expected in the base population. This approach, which also requires further development, might allow research funds to be allocated more judiciously.

**Data Analysis.** Many of the recent developments and ideas for new study designs and data collection that were discussed in this article require parallel developments in statistical methods. For example, the analyst might have to deal with complex sampling strategies (as in two-stage designs); missing, misclassified, and/or aggregated data on relevant variables; time-dependent

covariates; lag periods between first exposure and disease detection; incomplete case detection; and a limited sample size that severely restricts the number of covariates treated simultaneously. Several of these issues are covered further by Hatch and Thomas and by Prentice and Thomas in this issue. 

## REFERENCES

- Schlesselman JJ. Case-control studies: design, conduct, analysis. New York: Oxford University Press, 1982.
- Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research: principles and quantitative methods. Belmont, CA: Lifetime Learning Publications, 1982.
- Miettinen OS. Theoretical epidemiology: principles of occurrence research in medicine. New York: John Wiley & Sons, 1985.
- Kelsey JL, Thompson WD, Evans AS. Methods in observational epidemiology. New York: Oxford University Press, 1986.
- Rothman KJ. Modern epidemiology. Boston, MA: Little, Brown and Co., 1986.
- Checkoway H, Pearce N, Crawford-Brown DJ. Research methods in occupational epidemiology. New York: Oxford University Press, 1989.
- Leaverton PE, ed. Environmental epidemiology. New York: Praeger, 1982.
- Chiazze L Jr, Lundin FE, Watkins D, eds. Methods and issues in occupational and environmental epidemiology. Ann Arbor, MI: Ann Arbor Science Publishers, 1983.
- Goldsmith JR, ed. Environmental epidemiology: epidemiological investigation of community environmental health problems. Boca Raton, FL: CRC Press, 1986.
- Kopfler FC, Craun GF, eds. Environmental epidemiology. Chelsea, MI: Lewis Publishers, 1986.
- Poole C. Would vs should in the definition of secondary study base (letter). *J Clin Epidemiol* 43:1016-1017 (1990).
- Miettinen OS. The concept of secondary base (reply). *J Clin Epidemiol* 43:1017-1020 (1990).
- Morgenstern H, Kleinbaum DG, Kupper LL. Measures of disease incidence used in epidemiologic research. *Int J Epidemiol* 9:97-104 (1980).
- Breslow NE, Day NE. Statistical methods in cancer research, vol. 1. The analysis of case-control studies. Lyon: International Agency for Research on Cancer, 1980; 50-51.
- Rubin DB. Bayesian inference for causal effects: the role of randomization. *Ann Stat* 6:34-58 (1978).
- Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. *Int J Epidemiol* 15:412-418 (1986).
- Greenland S, Schlesselman JJ, Criqui MH. The fallacy of employing standardized regression coefficients and correlations as measures of effect. *Am J Epidemiol* 123:203-208 (1986).
- Greenland S, Maclure M, Schlesselman JJ, Poole C, Morgenstern H. Standardized regression coefficients: a further critique and review of some alternatives. *Epidemiology* (in press).
- Greenland S, Thomas DC. On the need for the rare disease assumption in case-control studies. *Am J Epidemiol* 116:547-553 (1982).
- Greenland S, Thomas DC, Morgenstern H. The rare-disease assumption revisited: a critique of estimators of relative risk for case-control studies. *Am J Epidemiol* 124:869-876 (1986).
- Armenian HK, Lilienfeld AM. Incubation period of disease. *Epidemiol Rev* 5:1-15 (1983).
- Thomas DC. Pitfalls in the analysis of exposure-time-response relationships. *J Chron Dis* 40 (Suppl 2):71S-78S (1987).
- Shy CM, Kleinbaum DG, Morgenstern H. The effect of misclassification of exposure status in epidemiological studies of air pollution health effects. *Bull N Y Acad Med* 54:1155-1165 (1978).
- Fleiss JL. Statistical factors in early detection of health effects. In: New and sensitive indicators of health impacts of environmental agents (Underhill DW, Radford EP, eds). Pittsburgh, PA: University of Pittsburgh, Center for Environmental Epidemiology, 1986; 9-16.
- Dosemeci M, Wacholder S, Lubin JH. Does nondifferential misclassification of exposure always bias a true effect toward the null value? *Am J Epidemiol* 132:746-748 (1990).
- Susser M, Stein Z. Third variable analysis: application to causal sequences among nutrient intake, maternal weight, birthweight, placental weight, and gestation. *Stat Med* 1:105-120 (1982).
- Cook TD, Campbell DT. Quasi-experimentation: design and analysis issues for field settings. Boston, MA: Houghton Mifflin Co., 1979.
- Greenland S. Randomization, statistics, and causal inference. *Epidemiology* 1:421-429 (1990).
- Buck C, Donner A. The design of controlled experiments in the evaluation of nontherapeutic interventions. *J Chron Dis* 35:531-538 (1982).
- Cornfield J. Randomization by group: a formal analysis. *Am J Epidemiol* 108:100-102 (1978).
- Donner A, Birkett N, Buck C. Randomization by cluster: sample size requirements and analysis. *Am J Epidemiol* 114:906-914 (1981).
- Zelen M. A new design for randomized clinical trials. *N Engl J Med* 300:1242-1245 (1979).
- Ast DB, Schlesinger ER. The conclusion of a 10-year study of water fluoridation. *Am J Public Health* 46:265-271 (1956).
- Greenland S, Morgenstern H. Classification schemes for epidemiologic research designs. *J Clin Epidemiol* 41:715-716 (1988).
- Greenland S. Response and follow-up bias in cohort studies. *Am J Epidemiol* 106:184-187 (1977).
- Robins J. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *J Chron Dis* 40(Suppl 2):139S-161S (1987).
- Robins J. The control of confounding by intermediate variables. *Stat Med* 8:679-701 (1989).
- Newman SC. Odds ratio estimation in a steady-state population. *J Clin Epidemiol* 41:59-65 (1988).
- Miettinen OS. The case-control study: valid selection of subjects. *J Chron Dis* 38:543-548 (1985).
- Greenland S, Morgenstern H. Matching and efficiency in cohort studies. *Am J Epidemiol* 131:151-159 (1990).
- Thompson WD, Kelsey JL, Walter SD. Cost and efficiency in the choice of matched and unmatched case-control study designs. *Am J Epidemiol* 116:840-851 (1982).
- Greenland S. Adjustment of risk ratios in case-base studies (hybrid epidemiologic designs). *Stat Med* 5:579-584 (1986).
- Mantel N. Synthetic retrospective studies and related topics. *Biometrics* 29:479-486 (1973).
- Kupper LL, McMichael AJ, Spirtas R. A hybrid epidemiologic study design useful in estimating relative risk. *J Am Stat Assoc* 70:524-528 (1975).
- Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 73:1-11 (1986).
- Miettinen OS, Wang JD. An alternative to the proportionate mortality ratio. *Am J Epidemiol* 114:144-148 (1981).
- Wang JD, Miettinen OS. The mortality odds ratio (MOR) in occupational mortality studies—selection of reference occupation(s) and reference cause(s) of death. *Ann Acad Med* 13(Suppl 2):312-316 (1984).
- Butler WJ, Park RM. Use of the logistic regression model for the analysis of proportionate mortality data. *Am J Epidemiol* 125:515-523 (1987).

49. Robins JM, Blevins D. Analysis of proportionate mortality data using logistic regression models. *Am J Epidemiol* 125:524-535 (1987).
50. Greenland S, Neutra R. An analysis of detection bias and proposed corrections in the study of estrogens and endometrial cancer. *J Chron Dis* 34:433-438 (1981).
51. Feinstein AR. Methodologic problems and standards in case-control research. *J Chron Dis* 32:35-41 (1979).
52. White JE. A two-stage design for the study of the relationship between a rare exposure and a rare disease. *Am J Epidemiol* 115:119-128 (1982).
53. Walker AM. Anamorphic analysis: sampling and estimation for covariate effects when both exposure and disease are known. *Biometrics* 38:1025-1032 (1982).
54. Cain K, Breslow NE. Logistic regression analysis and efficient design for two-stage studies. *Am J Epidemiol* 128:1198-1206 (1988).
55. Breslow NE, Zhao LP. Logistic regression for stratified case-control studies. *Biometrics* 44:891-899 (1988).
56. Flanders WD, Greenland S. Analytic methods for two-stage case-control studies and other stratified designs. *Stat Med* 10:739-747 (1991).
57. Louis TA, Lavori PW, Bailar JC III, Polansky M. Crossover and self-controlled designs in clinical research. *N Engl J Med* 310:24-31 (1984).
58. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. *Am J Epidemiol* 133:144-153 (1991).
59. Caporaso NE, Hayes RB, Dosemeci M, Hoover R, Ayesch R, Hetzel M, Idle J. Lung cancer risk, occupational exposure, and the debrisoquine metabolic phenotype. *Cancer Res* 49:3675-3679 (1989).
60. Mack W, Langholz B, Thomas DC. Survival models for familial aggregation of cancer. *Environ Health Perspect* 87:27-35 (1990).
61. Susser E, Susser M. Familial aggregation studies: a note on their epidemiologic properties. *Am J Epidemiol* 129:23-30 (1989).
62. Claus EB, Risch NJ, Thompson WD. Age at onset as an indicator of familial risk of breast cancer. *Am J Epidemiol* 131:961-972 (1990).
63. Morgenstern H. Uses of ecologic analysis in epidemiologic research. *Am J Public Health* 72:1336-1344 (1982).
64. Ohno Y, Aoki K, Aoki N. A test of significance for geographic clusters of disease. *Int J Epidemiol* 8:273-281 (1979).
65. Mollie A, Richardson S. Empirical Bayes estimates of cancer mortality rates using spatial models. *Stat Med* 10:95-112 (1991).
66. Mahoney MC, Labrie DS, Nascam PC, Wolfgang PE, Burnett WS. Population density and cancer mortality differentials in New York State, 1978-1982. *Int J Epidemiol* 19:483-490 (1990).
67. Lee JAH, Petersen GR, Stevens RG, Vesanen K. The influence of age, year of birth, and date on mortality from malignant melanoma in the populations of England and Wales, Canada, and the white population of the United States. *Am J Epidemiol* 110:734-739 (1979).
68. Lee JAH. Melanoma and exposure to sunlight. *Epidemiol Rev* 4:110-136 (1982).
69. Wallenstein S, Gould MS, Kleinman M. Use of the scan statistic to detect time-space clustering. *Am J Epidemiol* 130:1057-1064 (1989).
70. Roberson PK. Controlling for time-varying population distributions in disease clustering studies. *Am J Epidemiol* 132:S131-S135 (1990).
71. Hatch M, Susser M. Background gamma radiation and childhood cancers within 10 miles of a U.S. nuclear plant. *Int J Epidemiol* 19:546-552 (1990).
72. Ostrom CW Jr. Time series analysis: regression techniques, 2nd ed., quantitative applications in the social sciences, 07-009. Newbury Park, CA: Sage Publications, 1990.
73. McDowall D, McCleary R, Meidinger EE, Hay RA Jr. Interrupted time series analysis. Quantitative applications in the social sciences, 07-021. Newbury Park, CA: Sage Publications, 1980.
74. Sayrs LW. Pooled time series analysis. Quantitative applications in the social sciences, 07-070. Newbury Park, CA: Sage Publications, 1989.
75. Catalano R, Serxner S. Time series designs of potential interest to epidemiologists. *Am J Epidemiol* 126:724-731 (1987).
76. Gruchow HW, Rimm AA, Hoffmann RG. Alcohol consumption and ischemic heart disease mortality: Are time-series correlations meaningful? *Am J Epidemiol* 118:641-650 (1983).
77. Darby SC, Doll R. Fallout, radiation doses near Dounreay, and childhood leukaemia. *Br Med J* 294:603-607 (1987).
78. Crawford MD, Gardner MJ, Morris JN. Changes in water hardness and local death-rates. *Lancet* 2:327-329 (1971).
79. Goodman LA. Some alternatives to ecological correlation. *Am J Sociol* 64:610-625 (1959).
80. Beral V, Chilvers C, Fraser P. On the estimation of relative risk from vital statistical data. *J Epidemiol Community Health* 33:159-162 (1979).
81. Greenland S. Divergent biases in ecologic and individual-level studies. *Stat Med* 11:1209-1223 (1992).
82. Selvin HC. Durkheim's suicide and problems of empirical research. *Am J Sociol* 63:607-619 (1958).
83. Langbein LI, Lichtman AJ. Ecological inference, series 07-010. Beverly Hills, CA: Sage Publications, 1978.
84. Firebaugh G. A rule for inferring individual-level relationships from aggregate data. *Am Sociol Rev* 43:557-572 (1978).
85. Richardson S, Stucker I, Hémon D. Comparisons of relative risks obtained in ecological and individual studies: some methodological considerations. *Int J Epidemiol* 16:111-120 (1987).
86. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. *Int J Epidemiol* 18:269-274 (1989).
87. Greenland S, Morgenstern H. Neither within-region nor cross-regional independence of exposure and covariates prevents ecological bias (letter). *Int J Epidemiol* 20:816-817 (1991).
88. Robinson WS. Ecological correlations and the behavior of individuals. *Am Sociol Rev* 15:351-57 (1950).
89. Duncan OD, Cuzzort RP, Duncan B. Statistical geography: problems in analyzing areal data. Westport, CT: Greenwood Press, 1961.
90. Greenland S, Morgenstern H. Ecological bias and confounding (reply). *Int J Epidemiol* 19:766-767 (1990).
91. Richardson S, Hémon D. Ecological bias and confounding (letter). *Int J Epidemiol* 19:764-766 (1990).
92. Rosenbaum PR, Rubin DB. Difficulties with regression analyses of age-adjusted rates. *Biometrics* 40:437-443 (1984).
93. Brenner H, Savitz DA, Jöckel KH, Greenland S. The effects of non-differential exposure misclassification in ecological studies. *Am J Epidemiol* 135:85-95 (1992).
94. Greenland S. The effect of misclassification in the presence of covariates. *Am J Epidemiol* 112:564-569 (1990).
95. Savitz DA, Baron AE. Estimating and correcting for confounder misclassification. *Am J Epidemiol* 129:1062-1071 (1989).
96. Brenner H, Savitz DA, Greenland S. The effects of nondifferential confounder misclassification in ecologic studies. *Epidemiology* 3:456-459 (1992).
97. Stavratsky KM. The role of ecologic analysis in studies of the etiology of disease: a discussion with reference to large bowel cancer. *J Chron Dis* 29:435-444 (1976).
98. Connor MJ, Gillings D. An empiric study of ecological inference. *Am J Public Health* 74:555-559 (1984).
99. Polissar L. The effect of migration on comparison of disease rates in geographic studies in the United States. *Am J Epidemiol* 111:175-182 (1980).
100. Morgenstern H, Horwitz SM, Berkman LF. Connections between epidemiology and health services research: a review of psychosocial effects on childhood morbidity and pediatric medical care use. *J Ambulatory Care Management* 9:33-45 (1986).
101. Walter, SD, and Irwig, LM. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *J Clin Epidemiol* 41:923-937 (1988).
102. Marshall RJ. Validation study methods for estimating exposure proportions and odds ratios with misclassified data. *J Clin Epidemiol* 43:941-947 (1990).
103. Greenland S. Statistical uncertainty due to misclassification: implications for validation substudies. *J Clin Epidemiol* 41:1167-1174 (1988).
104. Spiegelman D, Gray R. Cost-efficient study designs for binary response data with Gaussian covariate measurement error. *Biometrics* 47:851-870 (1991).
105. Boyd LH Jr, Gudmund RI. Contextual analysis: concepts and statistical techniques. Belmont, CA: Wadsworth Publishing Co., 1979.
106. Lincoln JR, Zeitz G. Organizational properties from aggregate data: separating individual and structural effects. *Am Sociol Rev* 45:391-408 (1980).
107. Aitkin M, Longford N. Statistical modeling issues in school effectiveness studies. *J Roy Statist Soc (Series A)* 149(Part 1):1-43 (1986).
108. Iversen GR. Contextual analysis. Quantitative applications in the social sciences, 07-081. Newbury Park, CA: Sage Publications, 1991.
109. Humphreys K, Carr-Hill R. Area variations in health outcomes: artefact or ecology. *Int J Epidemiol* 20:251-258 (1991).
110. Khoury MJ, Beaty TH, Flanders WD. Epidemiologic approaches to the use of DNA markers in the search for disease susceptibility genes. *Epidemiol Rev* 12:41-55 (1990).
111. Morgenstern H, Winn DM. A method for determining the sampling ratio in epidemiologic studies. *Stat Med* 2:387-396 (1983).